Exploiting Vocabulary, Morphological, and Subtree Knowledge to Improve Chinese Syntactic Analysis(Abstract_要旨)

AUTHOR(S):

Shen, Mo

（ 続紙 1 ）

京都大学	博士（情報学）	氏名	沈　黙（Shen Mo）
論文題目	Exploiting Vocabulary, Morphological, and Subtree Knowledge to Improve Chinese Syntactic Analysis （語彙的、形態的、および部分木知識を用いた中国語構文解析の精度向上）		

（論文内容の要旨）

This dissertation investigates issues in various tasks of Chinese syntactic analysis, including word segmentation, part-of-speech (POS) tagging, and dependency parsing. To address the issues, the dissertation proposes a set of language resources and analysis methods exploiting vocabulary, morphological, and subtree knowledge, which are the contributions of this study. The dissertation consists of six chapters, which include four chapters describing the proposed resources and methods.

In Chapter 1, a discussion of the difficulties in Chinese word segmentation, POS tagging, and dependency parsing is presented. Several key observations are made in this chapter, which lead to the motivation of the approaches proposed in later chapters. An overview of the proposed approaches is also given.

In Chapter 2, based on the observations of the weakness of the existing morphology-based word definition in the Penn Chinese Treebank, a complete set of annotation guidelines for Chinese word segmentation and a tagset for POS tagging are proposed to overcome the inconsistency and data sparsity problems. A label set for dependency labeling is also proposed which is both compatible with the Universal Dependency annotation convention and consistent with the word segmentation and part-of-speech tagging guidelines proposed in the same chapter. By manually re-annotating the entire Penn Chinese Treebank 5.0 (CTB5), the advantages of the proposed annotation approach compared to existing ones is demonstrated.

In Chapter 3, an algorithm that extracts substrings as reliable word boundary indicators is proposed. These substrings significantly enhance the accuracy of word segmentation systems. The algorithm takes linear time in the average case, which is an advantage in processing large-scale raw texts. The algorithm is used in word segmentation as well as unknown word extraction, and in evaluation the proposed method outperforms existing comparable systems in both tasks.

In Chapter 4, the usefulness of character-level POS in the task of Chinese POS tagging is investigated. A tagset that is designed specifically for the task of character-level POS tagging is proposed. Based on this tagset, the entire CTB5 is manually augmented with an extra layer of annotation on each character. From the manually augmented corpus, a lexicon that contains the character-level POS information for all the characters in CTB5 is compiled. A morphological analysis

system is also proposed and implemented which can perform character-level POS tagging jointly with word segmentation and word-level POS tagging. In evaluation, it is demonstrated that by incorporating the character-level POS information, the accuracy of word-level POS tagging can be significantly improved.

In Chapter 5, a dependency parse reranking approach is proposed to address the problem of limited context in the existing graph-based dependency parsing models. A feature set that makes fully use of dependency grammar is explored to capture global information with less restriction in the structure and the size of the subtree context. It exhaustively explores a candidate parse tree for features from the most simple to the most expressive while it maintains the efficiency in terms of training and parsing time. Algorithms for selecting and extracting informative subtrees, eliminating redundancies in the search space of subtrees, and efficiently encoding features based on the extracted subtrees are proposed and implemented. In evaluation, a comprehensive set of experiments are conducted to demonstrate the effectiveness of the proposed reranking approach in the task of Chinese dependency parsing. In addition, an end-to-end evaluation that combines all the methods and resources proposed in this dissertation is performed to demonstrate their effectiveness and compatibility when combined in an integrated system.

In Chapter 6, the dissertation is concluded with a discussion of remaining issues and future work.

注) 論文内容の要旨と論文審査の結果の要旨は１頁を３８字×３６行で作成し，合わせて，３，０００字を標準とすること．
論文内容の要旨を英語で記入する場合は，４００～１，１００wordsで作成し
審査結果の要旨は日本語５００～２，０００字程度で作成すること．

（続紙 ２ ）

（論文審査の結果の要旨）

本論文は，中国語形態素・構文解析の精度向上を目的として，中国語の言語学的特徴に基づくツリーバンクアノテーション手法，語彙的知識の自動獲得手法，文字ごとの品詞に着目した形態素解析手法，および大域的な部分木知識を用いた構文解析手法について研究し，その成果をまとめたものである．得られた主要な成果は以下の通りである．

１．従来のツリーバンクアノテーション手法は，中国語の形態素理論に基づいて単語境界を決めるものであり，形態素の分類に曖昧性がある場合に，単語境界に関する一致性に欠けるという問題があった．さらに，既存手法は単語の定義に曖昧性があるために，実世界のテキストを解析する場合に未知語の比率が高くなり，解析精度が悪化するという問題があった．本論文では，これらの問題を解決するために，形態素間の構文構造に基づく新しい単語定義基準と，それと一致する品詞および構文構造のアノテーション基準を考案し，既存のツリーバンクであるPenn Chinese Treebank 5.0の全文を対象にアノテーションを実施した．構築したツリーバンクを用いた解析実験および機械翻訳実験によって，その有効性を確認している．さらに，これらのアノテーション基準および構築したツリーバンクを整備，公開しており，広く利用可能となっている．

２．従来の中国語単語分割手法は，未知語に対する単語分割精度が低いという問題があった．本論文では，大規模テキストから単語および複合語を効率的に抽出するアルゴリズムを考案し，抽出した大規模単語・複合語リストを形態素解析システムにおいて利用することによって，単語分割精度を向上させることに成功した．

３．従来の中国語品詞タグ付け手法においては，主に単語単位の情報が用いられており，文字単位の情報はほとんど利用されていなかった．本論文では，中国語文字が持つ品詞の特徴に着目し，既存のツリーバンクに文字単位の品詞情報を効率的に追加する手法，および文字単位の品詞情報を扱うことができる形態素解析システムの開発に成功した．さらに本論文では，既存ツリーバンクに追加した文字単位の品詞情報を辞書として整備，公開しており，広く利用可能となっている．

４．これまで広く利用されているグラフに基づく構文解析手法では，学習時に利用する部分木の特徴量が極めて局所的であるという問題があった．本論文は，リランキングに基づく構文解析の枠組みを用いることによって，大域的な部分木知識を特徴量として利用する高精度な構文解析システムを開発することに成功した．

よって，本論文は博士（情報学）の学位論文として価値あるものと認める．また，平成２８年２月１８日に実施した論文内容とそれに関連した試問の結果合格と認めた．

注)論文審査の結果の要旨の結句には、学位論文の審査についての認定を明記すること。
更に、試問の結果の要旨（例えば「平成　年　月　日論文内容とそれに関連した口頭試問を行った結果合格と認めた。」）を付け加えること。

Ｗｅｂでの即日公開を希望しない場合は、以下に公開可能とする日付を記入すること。

要旨公開可能日：　　年　　月　　日以降